The Effect of Visual Aids on Reading Numeric Data Tables

Yongfeng Ji, Charles Perin, and Miguel A. Nacenta

Fig. 1: Plain table (top left), table with zebra striping (top right), color encoding (bottom left), and bar encodings (bottom right).

Abstract—Data tables are one of the most common ways in which people encounter data. Although mostly built with text and numbers, data tables have a spatial layout and often exhibit visual elements meant to facilitate their reading. Surprisingly, there is an empirical knowledge gap on how people read tables and how different visual aids affect people's reading of tables. In this work, we seek to address this vacuum through a controlled study. We asked participants to repeatedly perform four different tasks with four table representation conditions (plain tables, tables with zebra striping, tables with cell background color encoding cell value, and tables with in-cell bars with lengths encoding cell value). We analyzed completion time, error rate, gaze-tracking data, mouse movement and participant preferences. We found that color and bar encodings help for finding maximum values. For a more complex task (comparison of proportional differences) color and bar helped less than zebra striping. We also characterize typical human behavior for the four tasks. These findings inform the design of tables and research directions for improving presentation of data in tabular form.

Index Terms—Data Table, Visual Encoding, Visual Aid, Gaze Analysis, Zebra, Data Bars, Tabular Representations.

1 Introduction

In the 1970s, tables were described as *"nothing more than a systematic arrangement of items of information"* [50]. The role of tables was solely to show numbers as a matrix of rows and columns [30]. It was believed that *"They are not recommended for communicating data to the general public; tables are most useful for fellow professionals. Both constructing and reading tables require skill of a high order."* [30]. Fifty years later tables are found in mass media and used by millions of people every day, including through spreadsheet software.

Regardless of pervasiveness and usefulness [2, 10, 16, 21, 33, 36], advice about tables focuses on reducing verbosity [17], "adding vitality" [41], conveying a simple message [13] or proper spacing and ordering [47]. Following this advice often yields featureless tables consisting mostly of space-arranged text or numbers, with few additional visuals (e.g., some horizontal lines—vertical lines are frowned upon).

We consider tables from the information visualization perspective, were the visual elements and aids of the representation can have large potential effects on reading performance. Within visual aids, we distinguish between *visual features* and *visual encodings*.

Visual features are design choices that affect the look of a table, without visually encoding data values. These include using a different typeface for table headers, left-aligning cell values, and customizing the size and color of cell borders. They are discussed often on the Internet and in advice articles in research journals (e.g., [13, 17, 21, 41, 46–48]). One particularly interesting visual feature is zebra: the shading of alternative rows of a table, presumably to facilitate the readers' horizontal gaze movements by avoiding inadvertently switching to a different row (also vertically, but vertical zebras are much less common).

Visual encodings are visual aids that, unlike visual features, represent data within the table graphically. For example, MS Excel supports encoding cell values with cell background color and with 'data bars' of corresponding length in the cell. While such visual encodings have only appeared relatively recently in mainstream software, they have been used in visualization for centuries (visit `https://aviz.fr/Bertifier/Bibliography` for an annotated bibliography).

Despite the ubiquity of tables, the potential effect of visual features on table reading (beyond aesthetic preference), and the long history of table encodings to facilitate table reading, there are almost no studies on how people read tables and on the effect of visual aids. We address this gap through a controlled study in which we tested four tasks: finding the name of the row with the largest value in a given column, the name of the column with the largest value in a given row, the column with the highest proportional difference between values in two rows, and the value at the intersection of a given row and column. We studied plain gridded tables, zebra, color encoding and bar encoding.

- *Yongfeng Ji is with University of Victoria. E-mail: yongfengji@uvic.ca.*
- *Charles Perin is with University of Victoria. E-mail: cperin@uvic.ca.*
- *Miguel A. Nacenta is with University of Victoria. E-mail: nacenta@uvic.ca.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

We found that, for all tasks, at least one visual aid produced a sizable performance improvement but it was not always the same aid. Color and Bar encodings were helpful to find maxima, but surprisingly not for the more complex proportional differences task, for which the Zebra helped. We attribute the benefits of Zebra to its usefulness in gaze navigation of the table structure, from our analysis of gaze movement.

2 Related Work

Wright and Fox [50] wrote in 1970 that *"[...] there have been very few experimental studies of the use of tabulated information, so that almost nothing is known about the effectiveness of alternative formats. Yet this lack of knowledge [...] is not commensurate with the size of the problem. Even small differences between tables may be of great practical importance when the information being displayed has to be used by a large number of people [...]"*. This is still true today. This section first presents visual features, then in-cell visual encodings.

2.1 Visual Features of Tables

There are many design and ergonomic elements that can be considered visual features (design choices that affect the look of the table): font family and size of the table body, row and column headers; presence or absence of borders, rulers and their stroke style; padding within cells and between rows and columns; vertical and horizontal alignment of cell content; alternate shading of rows (zebra); and so on. These visual features can be customized in table creation and presentation software, but have not been formally researched. Although there are many guidelines for creating effective tables (e.g., [17, 47]), these appear to be based on practical experience.

Zebra striping is a visual feature of particular interest because of its non-aesthetic purpose: to improve legibility [5, 48]. The familiar horizontal stripes of alternating darkness (see Figure 1, top right) might prevent accidentally skipping from one row to another. Horizontal stripes might also facilitate jumping to and from rows because they create two visually distinct groups (*vice versa* for columns in a vertically striped table). This might come at the cost of adding unnecessary clutter according to Tufte's data-ink ratio principle [45].

We only identified three empirical studies of zebra. Enders [14] found no effect of a horizontal zebra (in terms of time and accuracy) in 5/6 data retrieval and comparison tasks. The sixth, more difficult task, showed faster completion times with zebra. In two subsequent studies, Enders et al. [15] found a reduction in errors with zebra for similar (but time-limited) tasks, and then a preference for traditional zebra striping over plain, lined and other table styles. Lee et al. [27] compared the usefulness of tables without zebra, with horizontal zebra, and with vertical zebra and found that participants were slightly faster with the plain condition than with either zebra. However, the task was to detect the presence of a target symbol in a table otherwise full of distractors, a task that is not a standard one with numeric tables.

2.2 Visual Encodings in Tables

The history of tabular representation can be tracked back thousands of years to ancient Sumerian cuneiform tablets [10, 32, 33]. Creating visual encodings to enhance their readability is almost as old as tables themselves [43], and has taken multiple forms since, such as mapping numerical cell values to grayscale shading [29], black and white patterns [1, 8, 12] and bars [9].

Bertin presented a range of visual encodings for visual analysis of tables that were implemented on physical devices [4] and later translated to the web with *Bertifier* [38]. Visualization and Visual Analytics systems such as *The Table Lens* [39] and *Taggle* [18] also support encoding columns in tabular representations — simple ones like color, shading, length, and position, and more advanced ones like aggregations with box plots, histograms and UpSet [28]. Commercial software such as MS Excel support a smaller subset of these variables — typically color shading and bars, often called *data bars*. MS Excel also supports sparklines [46], word-sized graphics with typographic resolution [23] that can *"facilitate holistic recognition of patterns, trends, and outliers in multivariate sequences"* [6]. However, Sparklines do not represent a single value in a cell, but instead represent series of values as glyphs. Researchers have also introduced new types of encodings that are applicable to tables, such as FatFonts [37], a hybrid representation that integrates numeric representations with a visual mapping to the "amount of ink" by manipulating the digits' shapes. The usefulness of these encodings to read data tables is yet to be determined.

Color shading consists of encoding each cell's value in its background color on a color scale, similar to heat maps [49]. There are three main color schemes in visualization: Spectral, sequential, and diverging [7]. Much previous work studied how to encode data with color, a full review of which is beyond scope here. Diverging scales are used to represent quantitative data when there is a meaningful middle value such as 0 and spectral ones are used to represent nominal data [7]. Since we focus on positive quantitative values, we chose the simplest type of color scale: sequential. Many effective sequential color scales exist, such as the blue-white-red color scale [22, 35, 44], or scales relying on luminance changes [51]. We know that color shading is useful for product comparison tasks [40] and for comparison of distant values in scalar fields [31], especially when overlaid with digits [20].

Data **Bar** encodings are similar to bar charts — a very effective encoding for quantitative data [3, 25]. It is easy to retrieve information from the bars [11], especially when they are parallel and their bases are aligned [42]. Data bars in tables are automatically aligned when displayed within tables, although only in one chosen direction. Bars will have reduced benefits when aligned along the direction in which their lengths vary, the same way that it is more difficult to compare positions on unaligned scales than on aligned scales [11].

3 Goals and Research Questions

Our overarching goal is to understand whether and how visual aids affect how people read data tables. We pose five research questions:

RQ1 Do color and bar visual encodings help read data tables?
RQ2 Does the zebra visual feature help read data tables?
RQ3 Which is more useful: color/bar visual encodings or the zebra visual feature?
RQ4 Which of color or bar provides most benefit for reading tables?
RQ5 Does mouse movement correlate with the ability to read tables?

We answer these research questions by examining four conditions (Section 4.1) across four tasks (Section 4.2) against two performance measures (Section 4.7).

4 Methodology

We designed a controlled within-subject experiment testing four table representations for four tasks (with institutional ethics approval).

4.1 Data table Visual Representations (main condition)

To answer our research questions, we selected four representations that: 1) are widely used (to maximize results applicability); 2) are static (so that results apply to printed as well as on-screen tables); and, 3) are different (to cover a range of visual aids).

Plain. The *Plain* table (see Figure 1.A) is the control condition. The only visual features besides characters and digits are i) the horizontal and vertical grid lines, ii) a larger font weight for the headers, with a centered alignment for the column headers and a left alignment for the row headers; and iii) a right alignment for the numbers inside cells. These visual aids are present in all other conditions for comparability.

Zebra. The *Zebra* table (see Figure 1.B) uses a light gray color on the background of cells in alternate rows. This is a common visual feature that is invariable to changes in data values.

Color. The *Color* table (see Figure 1.D) encodes numerical values by coloring the background of a numerical cell in proportion to the value in the cell. It is commonly referred to as "conditional formatting". We chose to encode the values using one of MS Excel default color scales — from white for low values to red for large values. The Color condition is a visual encoding because the visual appearance of the table is affected by changes in data values.

Bar. The *Bar* table (see Figure 1.C) encodes numerical values by adding a left-aligned horizontal bar in the background of a numerical cell, with length proportional to the value in the cell. It is commonly referred to as "data bars". Like the Color condition, Bar is a visual encoding.

4.2 Tasks

To select tasks representative of real world use we started by collecting tasks evaluated in previous research that empirically investigated tables and other grid-based visualizations (e.g., [14,20,27,31]). We then short-listed a set of tasks that together were representative of the real world, included both low-level and higher-level as well as value-dependent and value-independent tasks. We settled on these four:

TVertMax. *Finding the name of the row that has the maximum value in a given column.* Participants were given the name (header) of a column and were asked to find the name of the row which had the largest value for that column. An example prompt was: *"Find the name of the Candy which has the highest 'Rating' value"*. This is a low-level task requiring navigation and multiple value comparisons.

THorMax. *Finding the name of the column that has the maximum value in a given row.* It is the transpose of TVertMax: participants were given the name of a row and had to provide the name of the column with the largest value for that row.

TDiff. *Finding the column, out of several possible columns, with the largest proportional difference between the values in two specific rows.* Participants were given the names of two rows and were asked to find for which of the last four columns their value difference was proportionally larger. Example prompt: *"For the candies called 'Pixie Sticks' and 'Life Savers', find in which column in the last four columns of the table they are most different (proportionally)"*. If the value in the column 'Sugar Percent' was 20 for 'Pixie Sticks' and 25 for 'Life Savers' (ratio: 1.25), but 40 and 80 for 'Price' (ratio: 2), then the correct answer is 'Price'. This is a higher-level task involving substantial navigation and multiple value comparisons.

TValue. *Retrieving a value given the names of a row and of a column.* Participants were given a row's and a column's names and found the value of the corresponding cell. Example prompt: *"For the Anime called 'Ultimate horror', please find its 'Watching' value."*). This is a low-level task that involves navigating the table and retrieving a value.

4.3 Datasets and Stimuli

We created eight table datasets: two per condition, one for a preliminary task and one for the rest. The datasets were based on four real data sets, but sanitized to avoid unnecessary noise in performance due to variations in familiarity, data ranges and number of attributes. We selected data sets that: 1) are about a topic and with language that are familiar to most people; 2) contain data that is not overly familiar (i.e., participants are unlikely to know in advance or expect particular values of particular attributes); and, 3) are plausible and realistic. The topics are candy nutritional values, animation series' ratings, cereal nutritional values and movie ratings. We curated the data sets so that they all contain 30 rows, each with a header of two words (e.g., "Reese Pieces" for the Candy dataset), and 11 columns, each with a header of one to three words (e.g., "Sugar Percent"). This results in tables with $30 \times 11 = 333$ cells. We further sanitized the row and column headers to avoid similarities and to prevent them from acting as visual landmarks (e.g., words that would be too distinctive). We rendered the tables to high-resolution bitmap images that cover the full screen of the monitor used in the experiment pixel-by-pixel without interpolation.

All cell values are numeric, with digits between 10 and 99 (2 significant digits, no decimals, no negative numbers). Although textual and categorical data are common in tables, in this study we focus on numerical data, which also offers the experimental advantage of keeping the dimensions and structure of the table the same for all conditions. We altered values based on specific tasks to normalize the difficulty of tasks trials and avoid edge cases. For instance, we altered the numbers to prevent consecutive trials from having the same answer. For TVertMax, the given column was never the first or last column, and the maximum value was never the first or last row. The tasks were of variable difficulty, with the second highest value between 1 and 15 units lower than the target. For THorMax, the given row was never the first three or the last three rows. Also of variable difficulty, THorMax tasks had second highest values between 1 and 39 units. For TValue, the given column was never the first or last two columns, and the given row was never the first three or the last three rows. The target response value was unique; otherwise, the value does not significantly influence the difficulty of this task. For TDiff, the given columns were always the last four and the given rows were non-contiguous with either 3 or 4 rows in between. We also forced values in each trial of TDiff to have similar sets of ratios and ensured that the answer ratios were consistently larger than 2, while the other ratios were clearly smaller than 2.

Because different columns often represent different types of values with different scales, we encode each column independently in the Color and Bar conditions. This is a common design choice; applying a global color or bar encoding often does not make sense because the differences between values in columns in the order of tens would not be perceivable if another column has values in the order of millions. For Color, we map the minimum value of a column to the first color on the scale (white) and the maximum value of the column to the last value on the scale (red). For Bar, we map the value 0 to a length of 0 and the maximum value of the column to the maximum length value (i.e. the cell width). We made this choice to avoid having bars of length 0 for values that would be larger than 0.

4.4 Procedure

Participants provided consent and filled a demographic questionnaire. Then, they went through a preliminary task that consisted of reading a table and describing it for maximum 60 seconds for each condition (see supplementary materials). This allowed participants to familiarize themselves with the different conditions.

Participants then completed four task blocks — one block per task. For each task block, the experimenter first demonstrated the task to the participant through toy examples and the participant could ask questions until they understood the task. Then, the participant carried out four trials of that task with each condition (16 trials per task). The first trial of each task-condition combination is considered training and excluded from the analysis.

Participants completed tasks in a fixed order (TVertmax, THormax, TValue, TDiff). They were assigned an order for the conditions (counterbalanced across participants using all possible orders) and that order was the same in each task block. For each condition, all participants saw the same dataset (e.g., the 'Candy' dataset for trials with Plain).

For a given trial, participants were first shown a white screen with the prompt. Upon pressing the space bar, the screen showed the corresponding table (t_0). When the participant found the answer, they pressed the space bar (t_1), which turned the screen blank, and after what they provided the answer verbally to the experimenter for recording. If a participant misunderstood a task (e.g., if they provided a numeric value instead of the expected name of a row) the experimenter marked the trial as invalid, the participant repeated the trial, and the repetition was marked as invalid too. Invalid trials are excluded from the analysis. At the end of each task block, participants ranked the conditions in terms of *preference*, *speed* and *accuracy*.

Participants were not allowed to use their hands or fingers on the screen, as this would have distorted the gaze location measurements. However, they could use a mouse to move a cursor around the screen if they so chose. This enables a more realistic interaction matching how some people interact with tables in real contexts, although it also introduces variance in the performance of the tasks. The experimenter also asked participants to keep a stable comfortable distance to the screen, without coming closer to the screen to see the table from closer. The study lasted approximately 1 hour.

4.5 Participants

We recruited 24 participants (aged 19 to 48, average 27, 14 female, 10 male). Participants had to be able to see a computer screen at a regular sitting distance ($\sim$ 80 cm) without glasses or contact lenses (due to gaze-tracker constraints), not have a visual disability, come across data tables in their regular activities, and not have photo-sensitive epilepsy. We discarded the data of three participants: two who had to sit very close to the screen to see the tables well, affecting eye tracking, and one who exceeded the allotted time by 30 minutes.

Table 1: Summary of hypothesis tests. Dark-green cells indicate strongly supported hypotheses (prob > 95%); light-green cells indicate a weaker (prob > 90%) or partial support (e.g., only one of two comparisons holds); and the pink cell indicates evidence against the hypothesis.

		TVertMax		THorMax		TDiff		TValue	
		Time τ	**Error ε**	**Time τ**	**Error ε**	**Time τ**	**Error ε**	**Time τ**	**Error ε**
H1	Color and Bar encodings help	$\tau_{Color} < \tau_{Plain}$ $\tau_{Bar} < \tau_{Plain}$	$\varepsilon_{Color} < \varepsilon_{Plain}$ $\varepsilon_{Bar} < \varepsilon_{Plain}$	$\tau_{Color} < \tau_{Plain}$	No Support	No Support	No Support	No Support	No Support
H2	Zebra helps	No Support	No Support	No Support	No Support	$\tau_{Zebra} < \tau_{Plain}$	No Support	$\tau_{Zebra} < \tau_{Plain}$	No Support
H3a	Color and Bar better than Zebra	$\tau_{Color} < \tau_{Zebra}$ $\tau_{Bar} < \tau_{Zebra}$	$\varepsilon_{Color} < \varepsilon_{Zebra}$ $\varepsilon_{Bar} < \varepsilon_{Zebra}$	$\tau_{Color} < \tau_{Zebra}$	No Support	$\tau_{Zebra} < \tau_{Color}$ $\tau_{Zebra} < \tau_{Bar}$	No Support	N/A	N/A
H3b	Zebra better than Color and Bar	N/A	N/A	N/A	N/A	N/A	N/A	$\tau_{Zebra} < \tau_{Color}$ $\tau_{Zebra} < \tau_{Bar}$	No Support
H4	Color faster than Bar (τ) Bar more accurate than Color (ε)	$\tau_{Color} < \tau_{Bar}$	No Support	$\tau_{Color} < \tau_{Bar}$	No Support	$\tau_{Color} < \tau_{Bar}$	No Support	No Support	No Support

4.6 Apparatus

Participants sat at a table in an office, in front of a 32 inch 4:3 display (3840×2160 resolution, 60Hz) with mouse and keyboard. They wore a light, untethered head-mounted gaze tracking device[1] that allowed them to move their head and body freely. The experimenter sat to the right of the participant with an additional mouse and a screen not readable by the participant. The experimenter used this setup to input participants' answers to the trials in the experimenter interface.

4.7 Measurements

Trial Completion Time (τ). The time from when the table becomes visible to when the participant presses the space bar after having found an answer ($t_1 - t_0$).

Error Rate (ε). The proportion of trials not answered correctly.

Gaze Tracking Data. The video from the participant's viewpoint, which includes the gaze points.

Mouse Movement (μ). The distance traveled by the mouse per trial.

Subjective Rankings. Participants ranked conditions from 1 to 4 in terms of *preference*, *speed* and *accuracy* for each task.

4.8 Hypotheses

The *a priori* hypotheses and analyses that match to **RQ1–RQ4** were pre-registered[2]. We use the $Measurement_{Technique}$ notation to simplify reading comparisons. For example, $\tau_{Bar} > \tau_{Plain}$ means that completion time with Bar is larger than with Plain for the task at hand. 4 research questions $\times$ 4 tasks (TVertmax, THormax, TDiff and TValue) $\times$ 2 measures (completion time: τ and error: ε) results in 32 hypotheses.

H1 (RQ1, all tasks): Participants will complete TVertmax, THormax, TDiff and TValue faster and with fewer errors with visual encodings than without, i.e.: $\tau_{Color} < \tau_{Plain}$, $\tau_{Bar} < \tau_{Plain}$, $\varepsilon_{Color} < \varepsilon_{Plain}$ and $\varepsilon_{Bar} < \varepsilon_{Plain}$. **Rationale:** Visual encodings help find target locations without reading digits.

H2 (RQ2, all tasks): Participants will complete all tasks faster and with fewer errors with Zebra than with Plain:, i.e.: $\tau_{Zebra} < \tau_{Plain}$ and $\varepsilon_{Zebra} < \varepsilon_{Plain}$. **Rationale:** Zebra helps people navigate tables and has been shown it can help [14, 15].

H3a (RQ3, TVertMax, THorMax, TDiff): Participants will complete TVertmax, THormax, and TDiff faster and with fewer errors with visual encodings than with Zebra, i.e.: $\tau_{Color} < \tau_{Zebra}$, $\tau_{Bar} < \tau_{Zebra}$, $\varepsilon_{Color} < \varepsilon_{Zebra}$ and $\varepsilon_{Bar} < \varepsilon_{Zebra}$. **Rationale:** Value comparison subtasks (possibly easier with encodings) are harder and more time consuming than visual navigation subtasks (made easier with Zebra).

H3b (RQ3, TValue): Participants will complete TValue faster and with fewer errors with Zebra than with encodings, i.e.: $\tau_{Zebra} < \tau_{Color}$ and $\varepsilon_{Zebra} < \varepsilon_{Bar}$. **Rationale:** TValue does not include value comparison subtasks (made easier with encodings), only visual navigation subtasks (made easier with Zebra).

H4 (RQ4, all tasks): Participants will complete all tasks faster with Color than with Bar, but with fewer errors with Bar than with Color, i.e.: $\tau_{Color} < \tau_{Bar}$ and $\varepsilon_{Bar} < \varepsilon_{Color}$. **Rationale:** Bar might slow down participants because of more visual interference in a cell (sharp transitions under numbers) than Color. Conversely, Bar could result in fewer errors because people are better at estimating and comparing variations in length than variations in color brightness/saturation (e.g., [11]).

4.9 Analysis Approach

To answer our research questions, we analyze the measurements in Section 4.7 both quantitatively and qualitatively.

QUANTITATIVE ANALYSIS - To answer hypotheses **H1–H4** that rely on ε and τ, we use a Bayesian statistical approach based on Markov Chain Monte Carlo (MCMC) simulations [19, 26]. τ (log-transformed) is modeled through a Student-t distribution (robust to outliers), with distribution average as function of condition and participant. When we report average times, they are back-transformed values of averages in the log-transformed domain. ε is modeled through a binomial distribution. Chance of error is a logistical function dependent on condition and participant. We substituted the pre-registered error model because participants made very few errors, rendering it inadequate for the collected data. Our corrected approach corresponds better with modeling of error recommended by Kruschke [26, p. 621]. All priors were suitably uninformative. Posterior predictions reasonably matched the data. Every MCMC simulation had chains with good mix and sufficient resolution (all $psrf < 1.05$, $ESS > 10{,}000$).

The quantitative analysis tests for **RQ5** (i.e., about μ) are variations of the models above which incorporate a mouse movement variable and answer three subquestions: 1) is μ different in different conditions? 2) Does μ affect τ? 3) Does μ affect ε? These analyses were not pre-registered, because we did not speculate *a priori* on μ questions. We encoded all trials where $\mu > 2000px$ as "mouse trial". The $2000px$ threshold corresponds to movements of half the horizontal size of the 4K screen and separates well trials with little activity from trials with mouse activity (only 14/1152 trials had $0 < \mu < 2000$). The μ analyses are correlational, since we did not force nor discourage mouse use. Scripts and data are provided in supplementary materials.

The scores obtained from the subjective ranking of conditions in each task are not tested statistically, we simply report average rankings.

QUALITATIVE ANALYSIS - The qualitative analysis of gaze data helps us depict participant strategies to complete tasks and provide possible explanations for performance differences between conditions. We opted for a qualitative analysis because the level of granularity of the analysis needs to adapt depending on the observed behavior (i.e., *a posteriori*).

We performed a systematic visual analysis of the trial video recordings captured by the glasses-mounted camera that were overlaid with gaze data by the Tobii gaze-tracker software. We only analyzed each participant's last two trials with each condition to remove transitory behaviors of participants familiarizing themselves the visual aids or developing their strategy. We lost the videos for one participant due to a recording error. The number of videos manually analyzed per task is thus $23(participants) \times 4(conditions) \times 2(repetitions) = 184$, but we

[1] https://www.tobiipro.cn/product-listing/tobii-pro-glasses-2/

[2] https://osf.io/b67xu?view_only=b9cc56507fc54ae399d0f468d53474ed

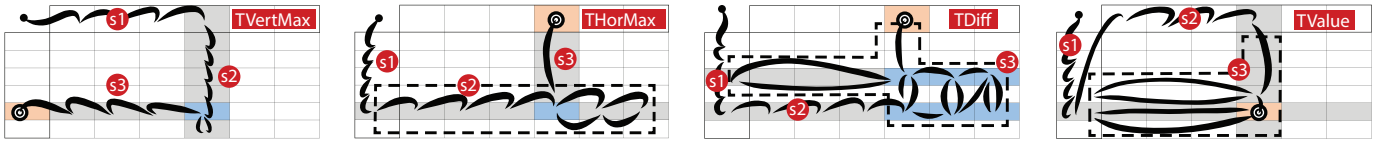

Fig. 2: Visual descriptions of predominant gaze patterns for each task. Labels s1, s2 and s3 refer to Step 1 to Step 3 we describe for each task.

could not analyze an additional 39 videos in TDiff (see Section 5.3). Our analysis process had 6 steps in total.

Steps 1 and 2. One author watched the videos with overlaid gaze positions, identifying types of movements (e.g., smooth scans, jumps), and their characteristics of interest (e.g., direction, presence of regressions). At each iteration, all the authors convened to validate the accuracy of the measurements and whether the categories were fit for purpose.

Steps 3 and 4. We tried to subcategorize the codes but realized that the large number of movement-characteristic combinations required instead an articulated language (see supplementary materials) to express types and sequences of movements. One author then used it to code all trials.

Step 5. The team verified that all movements were describable and validated a subset of the coded trials. The coded data is provided in supplementary materials.

Step 6. We queried the final coded video data for patterns that characterize the data (e.g., *Smooth Scans* are common when searching for a named column in the column headers) and differences in movements between conditions (e.g., *Jumps* are more common with Color than with Plain). These constitute the core of our qualitative analysis.

5 Results

Results are grouped by task. Table 1 summarizes hypothesis support, Figure 3 subjective rankings and Figure 2 dominant gaze patterns.

5.1 TVertMax Results

TVertMax is about finding the name of the row with the maximum value in a given column.

Completion Time and Error - Figure 4 shows that τ_{Color} is smallest (average 6.5 s) and is 89% smaller than τ_{Plain} (12.3 s), 90% smaller than τ_{Zebra} (12.4 s), and 16% smaller τ_{Bar} (7.6 s). This supports **H1**, (Color and Bar encodings help), **H3a** (Color and Bar are better than Zebra), and **H4** ($\tau_{Color} < \tau_{Bar}$), but not **H2**, since τ_{Zebra} is not smaller than τ_{Plain}.

Error levels are low (18 errors in total, or 6.2%, see Figure 5). Participants were less likely to make errors with Color than with Plain and Zebra (strong evidence) and with Bar than with plain and Zebra (slightly weaker evidence). This supports **H1** and **H3a**. **H2** is not supported since we did not find conclusive evidence that ε_{Zebra} is smaller than the other visual aids. **H4** is not supported in terms of error because there is no conclusive difference between ε_{Color} and ε_{Bar}.

Effects of Mouse Use - Participants were more likely to use the mouse with Plain and Zebra (40.3% and 41.7% of the time, respectively), than with Color and Bar (32.4% and 34.7%). Trials where the mouse was used in Zebra took almost certainly longer (> 99% probability), by an estimated 3.5 seconds on average; for the other techniques the differences are not conclusive. There is also no evidence that using the mouse affected the number of errors in any of the conditions.

Task Gaze Pattern - Most trials exhibit a clear three-step strategy illustrated in Figure 2. Participants completed these steps with consistent types of movement across all conditions.

Step 1. *Finding the column header indicated in the prompt.* 60% of the trials started with this step (in 40% the participants did not need to check the column header to find the required column). To find the column, participants went left to right, and mostly with Jumps (47/108, 43%) or Smooth Scans (62/108, 57%), but no Super Smooth Scans.

Step 2. *Finding the cell within the corresponding column that has the maximum value.* 99% of the analyzed trials contained this step. This step required slower movements, with no Jumps observed. Participants mostly used Jump Scans, in 75/183 trials without the mouse and in 28/183 trials with the mouse, and all with regressions. They also used Smooth Scans (25/183) and Super Smooth Scans (47/183), both split by about half on their use of the mouse, and all with regressions.

Step 3. *Finding the row header that corresponds to the cell with maximum value.* 99% of the analyzed trials contained this step. Participants mostly used right-to-left Smooth Scans without the mouse (89/183 trials, only 5 of these with regression) and with the mouse (42/183 trials, only 4 with regression). Jumps and jump scans were unusual (21 and 22 out of 183 respectively), with very few regressions or mouse use (6/183). As in Step 1, no participant used Super Smooth Scan.

Gaze Differences Between Conditions - We found some stark differences in gaze movement types between conditions:

- Participants used two to three times as many Jump Scans with encodings (28 for Color and 29 for Bar) compared to without encodings (10 for Plain and 10 for Zebra).
- Participants almost never used Super-Smooth Scans with encodings (0 for Color and 2 for Bar) but used them often without encodings (25 for Plain and 25 for Zebra).
- Participants used Jumps (Jump, Jump Scan and Mouse Jump combined) fewer times with Plain (19) than with Zebra (29), Color (44) and Bar (44).

Step-specific analysis reveals further condition differences in Step 2, where participants used Jump Scan (without mouse) more often with encodings (29 times with Color and 31 times with Bar) than with Plain (5) and Zebra (10). This was also the case for Jump Scan with mouse (Color: 13, Bar: 12, Plain: 2, Zebra: 1). Conversely, participants used Smooth Scan (without mouse) more with Plain (8) and Zebra (8) than with Color (2) and Bar (1) and Super-Smooth Scan (without mouse) more with Plain (11) and Zebra (10) than with Color (0) and Bar (1). They also used Smooth and Super-Smooth Scan with mouse more with Plain (16) and Zebra (14) than with Color (0) and Bar (2).

Subjective Results - Bar was ranked best, then Color, then Zebra then Plain, for all measurements (see Figure 3).

TVertMax Discussion - The results for TVertMax generally align with our expectation that Color and Bar help (with both time and error–**H1**). We can explain differences between conditions through gaze behavior: Color and Bar allowed participants to jump to candidate maximum values in a column, presumably because they could identify the most likely cells before fixating on number values. With Plain and Zebra, participants instead scanned vertically and read values in the column one by one. In other words, Step 2 is the step that takes the most time, and it is where these visual encodings help.

We can also link the slower completion times with Zebra and Plain to increased mouse usage. Using the mouse as bookmark or moving guide to keep gaze movements horizontal might have some benefits, but these seem overridden by the cost of controlling the mouse itself, hence the slower gaze movements with Zebra and Plain. Note that this analysis is correlational and mouse use can be cause or consequence.

Participants did not do better with Zebra than with Plain (**H2** not supported). We did not observe obvious differences in the types of movements when gaze had to move left to find the row header in Step 3, even though we expected Zebra to allow participants to move faster horizontally by reducing their fear of inadvertently "changing lanes".

We also found a clear completion time difference between Color and Bar (∼1 second). We speculate that, when not in the fovea, differences in color saturation are more salient than differences in length, allowing people to move more efficiently to candidate values.

		TVertMax	THorMax	TDiff	TValue
Plain	Preference	3.8	3.7	3.7	2.8
	Accuracy	3.8	3.7	3.7	3.0
	Speed	3.8	3.7	3.7	2.7
Zebra	Preference	2.9	2.4	1.7	1.3
	Accuracy	2.8	2.5	1.8	1.3
	Speed	3.0	2.5	1.5	1.4
Color	Preference	2.0	1.6	2.4	2.9
	Accuracy	2.1	1.6	2.4	2.9
	Speed	1.9	1.7	2.4	3.0
Bar	Preference	1.3	2.3	2.3	2.9
	Accuracy	1.3	2.3	2.1	2.9
	Speed	1.3	2.1	2.4	2.9

Fig. 3: Average subjective ranking (between 1 and 4) in terms of preference, accuracy and speed for the four conditions, per task. Darker cells (and smaller value) mean higher (better) ranks.

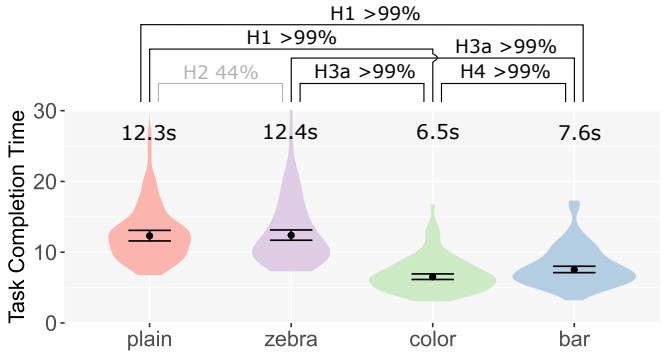

Fig. 4: Completion time for TVertMax. The widths of Violin plots indicate density of measurements. Error bars are 95% High-Density Intervals of the log-untransformed mean estimation (in seconds). Pairwise comparisons (lines at the top) are black if the estimated probability of a condition is $> 95\%$ or $< 5\%$ and gray otherwise.

TVertMax	Plain	Zebra	Color	Bar
Plain		53.0	**4.6**	7.5
Zebra	47.0		**3.4**	5.5
Color	**95.4**	**96.6**		69.5
Bar	92.5	94.5	30.5	

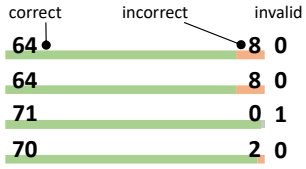

Fig. 5: Errors for TVertMax. Pairwise comparisons cells (test of column condition being more likely to produce errors than row condition, in percentages) are black if the estimated probability of a condition (number) is $> 95\%$ or $< 5\%$ (the conjugate), grey if $> 90\%$ or $< 10\%$ and white otherwise. The horizontal stacked bars on the right indicate the number of correct, incorrect, and invalid trials per condition (row).

5.2 THormax Results

THorMax is about finding the name of the column that has the maximum value in a given row.

Completion Time and Error - Figure 6 shows that τ_{Color} is shortest (8.9 s, over 9%, 11% and 12% shorter on average than $\tau_{Zebra} =$ 9.8 s, $\tau_{Bar} = 9.9$ s and $\tau_{Plain} = 10$ s respectively). Because only Color helped (not Bar), **H1** is only partially supported. **H2** is not supported

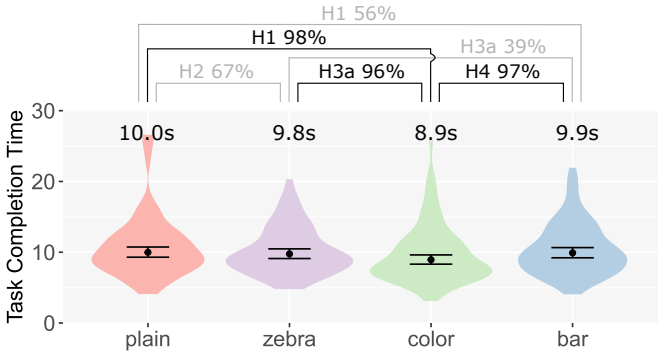

Fig. 6: THorMax completion time. Refer also to Figure 4 caption.

(no evidence that Zebra shortens time), **H3a** is only partially supported (only $\tau_{Color} < \tau_{Zebra}$) and **H4** is supported ($\tau_{Color} < \tau_{Bar}$).

Error levels (ε) are low (15 errors in total, or 5.2% of valid trials—see supplementary materials). ε_{Color} is largest, but statistical reliability is relatively weak (89% probability of Color being more error-prone than Bar, and 86% with respect to Plain and Zebra). No hypotheses are supported by the analysis of ε.

Effects of Mouse Use - Participants use the mouse more with Plain than with the other conditions (52.7% vs $<$ 44% in all others, all with probability $> 96\%$). We also found that using the mouse is associated with smaller τ values. When participants used the mouse, this added 2.9 s on average to τ_{Color}, 2.5 s to τ_{Plain}, 1.2 s to τ_{Zebra} and 1.0 s to τ_{Bar}). The evidence does not support a ε-μ correlation.

Task Gaze Pattern - Most trials exhibit a clear three-step pattern illustrated in Figure 2 — the transposed version of that in TVertMax.
Step 1. *Finding the row header indicated in the prompt.* 98% of the trials started with this step. Most participants used vertical Smooth Scan (in 172/178 trials), with mouse in 89/172 trials and with regression in 65/172 trials. There were no Jumps, very few Jump Scans (6/178), and no Super-Smooth Scans.
Step 2. *Finding the cell within the corresponding row that has the maximum value.* Like in Step 1, most participants used Smooth Scan (in 140/184 trials) – all but two with regression. The second preferred gaze movement was Super-Smooth Scan (in 47/184 trials), with mouse in about half the trials. All Super-Smooth Scans were with regression. There were 0 Jumps and 0 Jump Scans.
Step 3. *Finding the column header that corresponds to the cell with maximum value.* Most participants used a bottom to top Jump, without regression, and without mouse (in 145/184 trials). Second is Smooth Scan, again without regression and without mouse (in 18/184 trials).

Gaze Differences Between Conditions - The analysis of gaze patterns did not reveal any clear differences between conditions, either globally or when analyzing based on the different steps.

Subjective Results - Color was ranked best, then Bar, then Zebra then Plain, for all measurements (see Figure 3).

THorMax Discussion - We included THorMax as a counterpoint to TVertMax so that we could understand the symmetry or asymmetry of effects due to the anisotropy of tables (cells are wider than tall, there are more rows than columns). However, unlike TVertMax, THorMax is not a natural task for most tables because columns, which usually represent attributes, have wildly varying measurement units and orders of magnitude (e.g., one column could count milligrams of potassium — 0.01 to 0.2, and the next the percentage of sugar — 40% to 80%). This means that, unless columns happen to be homogeneous (e.g., the same measurement for a series of years), it is usually meaningless to find the attribute with the largest value within a row. As a consequence, value encodings make more sense *per column*, which means that comparing encodings *per row* can lead to errors when a value is the maximum in a column, yet low in the row. This was the case in our stimuli and it explains the higher number of errors with Color, because we

found that all errors belonged to the group of trials where the encoding "contradicts" the values (7 out of 16 trial types).

Despite this, most participants carried out the task correctly in all conditions, and faster with Color – which we explain in two ways. One is that the (partially misleading) information provided by the encoding was still leveraged by participants to find the highest value candidates faster, in the same way as in TVertMax. An alternative is that the background color of the cells created useful landmarks when traversing and comparing the values in the row. We lean towards the latter because most of the gaze movements in Step 2 were Smooth Scans and Super-Smooth Scans with regression, while using color information to identify candidates would have resulted in more Jumps or Jump Scans.

Bar did not help, likely because it is difficult to compare lengths when bars are aligned in the direction that they vary (horizontally in our case). This makes Bar ineffective, in stark contrast with what happened in TVertMax. This is not the only consequence of the anisotropy of tables: in Step 3, to reach the column header, the dominant movement was Jump; while in TVertMax, Step 3 consisted mostly of Smooth Scans. In other words, it seems that participants naturally recognized that moving horizontally along a narrow row to its beginning (Step 3 in TVertMax) is much riskier than jumping vertically to the wider (and less distant) column header, and adapted accordingly.

As in TVertMax, participants did not do better with Zebra than with Plain (**H2** not supported). The potential benefit of Zebra helping "stay on the lane" when following a row to find its maximum (Step 2) might be negligible when the gaze movement is a non-risky smooth scan.

5.3 TDiff Results

TDiff is about finding the column, out of several possible columns, that has the largest proportional difference between two given rows.

Completion Time and Error - As expected, TDiff being more complex than the other tasks, it took longer to complete (around half a minute in all conditions). Figure 7 shows that τ_{Zebra} is smallest (30.46 s average) and is 5.8% smaller than τ_{Color} ($\sim$ 2 s), 9.5% smaller than τ_{Plain} ($\sim$ 3 s) and 12% smaller than τ_{Bar} ($\sim$ 4 s). We found no conclusive evidence that these encodings help (**H1**); **H2** is clearly supported (Zebra helps); **H3a** is contradicted by the results, with $\tau_{Zebra} < \tau_{Bar}$ and $\tau_{Zebra} < \tau_{Color}$; **H4** is moderately supported, with $\tau_{Color} < \tau_{Bar}$.

Participants made more errors in this task (17% of trials—See supplementary materials), but we see no clear differences between conditions.

Effects of Mouse Use - Participants used the mouse in most trials (91%), similarly across all conditions. There is also no conclusive evidence that using the mouse affects τ or ε.

Task Gaze Pattern - Of the 184 trials, we lost the videos for 24 trials from 3 participants when the eye-tracking device failed to record these trials, or when the trials exceeded the device's recording time limit. 15 other trials were too noisy for analysis. We analyze the 145 remaining trials. TDiff is a more complex task that can be performed in many different ways, and we did not identify clear strategies to complete it. However, we did identify two common initial steps (Step 1, Step 2) followed by a group of diverse and somewhat chaotic actions with lots of variations between participants (Step 3). Figure 2 shows the typical gaze patterns for TDiff.

Step 1. *Finding the two row headers indicated in the prompt.* 99% of the trials started with this step, and 42% came back to this step after they went through Step 2. Most participants used Smooth Scan (in 143/145 trials), with or without mouse. Most trials were with mouse (124/145). This step included regression in most trials without mouse (91%) and less often in those with mouse (65%). There were no Jumps, Jump Scans, or Super-Smooth Scans.

Step 2. *Finding the horizontal location of the values to compare by identifying the corresponding columns based on the column header.* Most participants used Smooth Scan (in 113/145 trials). Of those, 63.5% were with mouse, with regression (54/145 trials) or without (37/145 without). Participants who used the mouse tended to hover the cursor over the second row of interest, then kept gaze and mouse position vertically synchronized. There were no Jumps or Jump Scans.

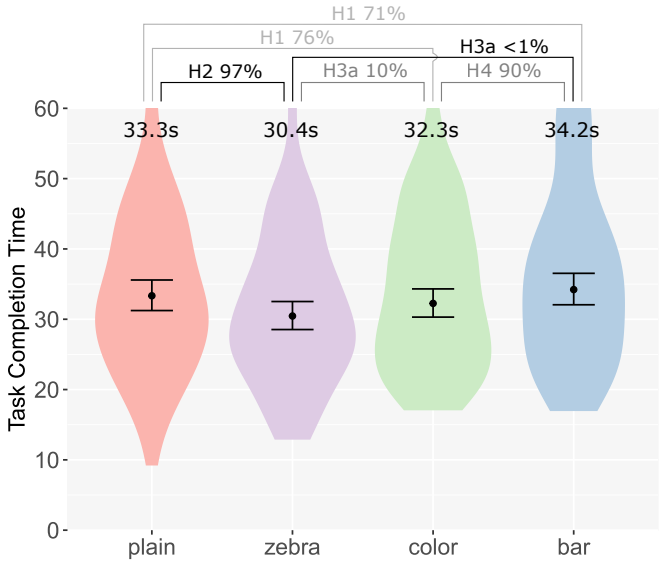

Fig. 7: TDiff completion time. Refer also to Figure 4 caption.

Step 3 (group). *Comparing Cells.* After completing Step 2, 42% of trials (61/145 trials) included a horizontal back-and-forth Jump to the row header, without mouse and without regression. Presumably, this was to check that the viewer was still looking at the right row. After that, different types of movement patterns took place. For example, in a *Compare* pattern, participants moved up and down between two relevant cells in the same column several times. Another common pattern was a jump up and back to column headers or additional jumps left and back to row headers between instances of 'Compare', presumably to check they were still comparing the correct rows and columns.

Gaze Differences Between Conditions - The analysis of gaze patterns did not reveal any clear differences between conditions, either globally or when analyzing based on the different steps.

Subjective Results - Zebra was ranked best, then Bar, then Color then Plain, for all measurements (see Figure 3).

TDiff Discussion - TDiff is likely the most complex task studied empirically on tables to date. We expected Color and Bar to be helpful because we assumed that visually decoding values would be quicker than decoding the symbolic digits. We did not find evidence of this (**H1** not supported). Instead, only Zebra seems to help (**H2** is supported). As a consequence, **H3a** is also not supported (the benefits of these encodings are negligible compared to the benefits of Zebra).

The evidence points to three possible insights: A) Color and Bar might not help (or help too little) when estimating the proportional difference between two cells. This might be because the estimation task is easier to do with digits than with bars or shaded cells, or because the setup of the task hinders this operation (e.g., the estimation requires comparison of cells across some space, which is filled by other encoded cells which interfere). B) Navigational sub-actions required by the task, which are presumably helped by Zebra, dominate the completion time for the task. C) Landmarks created by these encodings might not help much with navigation, at least compared to Zebra.

We also suspect that participants' time to decide on a task-completion strategy affects the timing and hides effects that would be more prominent after participants have settled on that strategy.

5.4 TValue Results

TValue is about retrieving the value contained in the cell specified by the names of its row and column.

Completion Time and Error - Figure 8 shows that τ_{Zebra} is smallest (9.5 s, 9%, 14% and 16% smaller than $\tau_{Color} = 10.3$ s, $\tau_{Bar} = 10.8$ s and $\tau_{Plain} = 11$ s, respectively). There is no statistically reliable

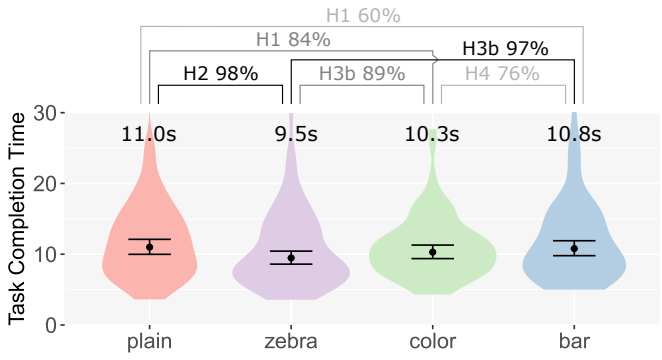

Fig. 8: TValue completion time. Refer also to Figure 4 caption.

difference between τ_{Plain} and τ_{Bar} nor between τ_{Plain} and τ_{Color} (**H1** not supported). **H2** is clearly supported ($\tau_{Zebra} < \tau_{Plain}$). **H3b**, which is specific to this task, is supported, although the evidence that $\tau_{Zebra} < \tau_{Bar}$ is much stronger than that for for $\tau_{Zebra} < \tau_{Color}$.

There were only six errors for TValue (see supplementary materials) — too small a number to infer any differences between conditions.

Effects of Mouse Use - Participants used the mouse in 70-76% of the trials, depending on the condition. There is no evidence that they were more likely to use the mouse in one condition than another. There is reliable evidence (98% probability) that mouse use reduced τ_{Bar} by about 1.7 seconds. There is no such evidence for the other conditions, and no evidence that mouse use correlates with error.

Task Gaze Pattern - Participants completed TValue in two variants, each decomposed into three steps, regardless of condition (see Figure 2). Only two trials started directly at step 3.
Steps 1 and 2. *Finding the item named in the prompt in header (step 1), then the other header (step 2).* 79% of the time, participants started by finding the row header in step 1, then the column header in step 2. Most participants used Smooth Scan in both steps (in 166/184 trials). We did not find an explicit Step 2 in 62 trials, presumably because participants remembered the location of the column or row. Participants were more likely to use the mouse while scanning the row header (103/164 trials were with mouse, i.e. 63% of trials) compared to the column header (10/138 trials were with mouse, i.e. 7% of trials). Also, 111/182 trials (61%) were with mouse for Step 1, against 2/120 (1.5%) for Step 2.
Step 3 (group). *Find the cell within the corresponding row/column that aligns with the specified column/row header* (depending on variant). In the first variant, participants traverse vertically the column that they found in step 2, with occasional Jumps and Scans to the row headers as they approach the target row, until they locate the crossing of row and column. In the second variant, they traverse horizontally the row found in step 2, with occasional Jumps and Scans to the column headers as they approach the target column. The predominant movements are Jumps (in 179/184 trials) and Smooth Scans (175/184). Almost all trials contained both, although Jumps are more than twice as numerous (737 Jumps out of the 1003 movements) than Smooth Scans (318/1003).

Gaze Differences Between Conditions - The analysis of gaze patterns did not reveal any clear differences between conditions, either globally or when analyzing based on the different steps.

Subjective Results - Zebra was ranked best, and Plain, Color and Bar received similar rankings, for all measurements (see Figure 3).

TValue Discussion - We initially thought TValue would be the simplest and most straightforward task. However, TValue requires a relatively large number of operations (e.g., going to, and back from, headers). This can be explained with the limitations of the human visual system: gaze locations are not always easy to return to, especially when there is high density of possible locations or when there are no distinctive landmarks confirming return to the appropriate location.

It is precisely this difficulty that might explain Zebra's advantage to complete TValue (**H2** supported). Vertically, the Zebra stripes reduce the density of elements to come back to by a factor of two (i.e., participants remember that they are traveling back to a dark or light row). Horizontally, the stripes may help participants "stay on the lane" when facilitating jumps back and forth to the headers.

As expected, Color and Bar did not help, since the cell value is irrelevant to this task (**H3b** supported). However, Color and Bar do produce visual landmarks that could have helped participants recover the gaze position in the table when jumping or scanning back and forth; however, the data did not show evidence of this.

This task also shows how the mouse is sometimes used as a "mobile landmark". Participants who started by scanning the row headers left the cursor at the appropriate row so that, when finding the correct column, they could more easily find the intersecting cell.

6 Discussion

Summarizing discussion from Section 5.1, Section 5.2, Section 5.3 and Section 5.4 provides answers to the research questions of Section 3: visual encodings help read data tables, at least for TVertMax and THorMax (**RQ1**); Zebra helps read data tables faster, at least for TDiff and TValue (**RQ2**); both visual encodings and visual features are useful, but in disjoint task sets (**RQ3**); the color encoding brings more benefits than the bar encoding (**RQ4**); and using the mouse seems to be detrimental unless it is used as a mobile landmark in complex tasks that otherwise require significant spatial memory such as TDiff and TValue (**RQ5**).

We further interpret findings around four themes: i) the nature of table tasks, ii) the benefits of visual aids; iii) the effect of mouse use, and iv) the shape of tables. **Design recommendations are in boldface**.

6.1 The Nature of Table Tasks and their Operations

Explaining differences between tasks and between conditions requires further distinctions in the operations required to perform each task. Our gaze analysis uncovered types of relevant operations — note that an *operation* does not necessarily correspond one-to-one to a *step* from our analysis. Some tasks rely heavily on *(gaze) navigational* operations. For example, TValue requires travelling left from a cell to check the name of its row header. Navigational operations tend to be fast, but their contribution to completion time is not negligible. Other operations are mostly *cognitive*, such as in Step 2 of TVertMax, which requires comparing cells in a column and remembering the highest value until the full column is scanned. Cognitive tasks such as scanning row or column headers to find a name or comparing two cell values to derive a proportion (in TDiff) still have substantial navigational components, and tend to be accompanied by regressions (repeated back and forth gaze movements between cells). In contrast, purely navigational operations do not have regressions. Finally, some navigational operations require spatial memory, such as returning to a previously seen cell.

Predictably, some tasks take longer to complete than others because they involve different operations. Unsurprisingly, the most sophisticated task (TDiff) takes about three times as long to complete as the other tasks (30 s vs. 10 s). Unexpectedly, participants took longer to complete TValue than THorMax and than TVertMax with encodings. This is explainable by the number and types of operations required. TValue, perhaps the most fundamental task in tables, involves a relatively large number of varied operations: scanning both headings sequentially to find the selected column and row (both cognitive operations with per-element matching), then jumping back to an approximate location (navigational with spatial memory), then one or more navigational jumps back and forth to the heading to verify the right location. In contrast, the flow of operations to find a maximum is much smoother: a scan of the header, then a linear navigation until the maximum value is found, and then a jump to the other header. Although finding the maximum theoretically involves a larger number of comparisons (e.g., each value to the memory-held maximum), the spatial memory-reliant completion of TValue results in many navigational operations that add up to a similar time. When finding a maximum is facilitated by encodings, this makes a task like TVertMax much faster than TValue. From here we derive a first — trivial yet important — recommendation: **when designing a table, consider the main tasks it will support**.

6.2 The Benefits of Visual Encodings and Visual Features

Because we found that no single aid offered advantages in all tasks, **selection of visual aids should be matched to the main tasks to be supported**. However, there were no cases of participants being more performant with Plain than with any visual aid (with the exception of the errors for Color in THorMax that are more an issue of communicating that encoding is applied by column and not by row). Knowing that visual aids do help, the question is which one(s) to use and when.

The encodings (Color and Bar) show important speed advantages for TVertMax, most likely because the cognitive subtask of reading each value in a column is replaced by reading a much smaller selection of candidate values to which it is easy to direct gaze with a saccade. Color is better than Bar, most likely because it is easier to perceive from further distances when not foveated. Therefore we recommend to **leverage Color to find extrema**, as long as the encoding is consistent with the task. This is consistent with findings from similar tasks tested in numeric scalar data fields [20, 31].

Bar also helps with TVertMax, but less. In addition, Bar not helping for THorMax echoes previous findings that there are differences in accuracy between aligned length/position and length [11]; but our results, for the tasks we tested, go one step further because Bar did not help *at all* when the bars are on an unaligned scale. Vertical bars would likely help for horizontal comparisons, but differences would be harder to see due to the typical aspect ratio of table cells (wide). Nevertheless, we speculate that Bar may have advantages, such as better compatibility with Zebra, and supporting spatial graphical thinking (e.g., easier to calculate midpoints or averages across cells).

Surprisingly, the encodings did not help with the more complex task we tested (TDiff), although we know an encoding like Bar is particularly suitable for proportion estimation [11]. It is troublesome that that encodings help perform simple, granular tasks, but might not help with complex, compound tasks. Encodings may be difficult to use for anything other than identifying large or small values when the intermediate space is filled with other values and their benefit might be constrained to cognitively cheap filtering, not to simplification of cognitive comparisons. Therefore, we recommend to **not assume that the advantages of visual aids in simple tasks will translate to larger tasks**, even for tasks with similar low-level operations.

For TDiff, only Zebra showed an advantage. This is likely because Zebra helped with the numerous vertical navigational operations required in this task, as participants could remember to focus on a darkened or clear background. Zebra also saved time in the TValue task, which requires similar navigational operations (see Figure 2). This is consistent with Enders' findings that horizontal Zebra striping is beneficial [14, 15], although only for their most difficult question. There might be operations beyond finding extrema that benefit from visual encodings and are substantial constituents of complex tasks. However, for the complex task we tested, we recommend to **use Zebra striping unless operations that benefit from visual encodings, such as finding extrema, constitute a large proportion of the task**. The benefits are, however, only around 10%.

Zebra's lack of benefit in TVertMax and THorMax indicates that it does not necessarily support horizontal visual flow [27] or "staying in the lane", but reduces vertical bandwidth to enable more confident vertical gaze navigation (it facilitates saccades to a dark or clear row). Further empirical research should confirm this insight; nevertheless **we should not assume that the benefits of the Zebra are exclusively in horizontal navigation**.

Finally, although it seemed plausible that encodings could help navigational tasks by creating visual landmarks through data patterns, we did not find any substantial evidence of this. This is most evident from the TValue task, where the encoding does not have any cognitive function and encodings are not better than Plain.

6.3 The Effect of Mouse Use

We observed that participants used the mouse in two ways: A) to accompany their gaze with the cursor (e.g., when scanning a heading) to improve their performance, perhaps recognizing that gaze is somewhat unreliable when scanning; and B) as a "bookmark" to keep track of a row or a column when the task required them to look away (e.g., when both the location of the row and the column is necessary).

Although the evidence is correlational (participants could choose) and partial (only for some conditions), the benefits of using the mouse to accompany gaze (A) do not seem to overcome the extra cost of moving the mouse (use of the mouse is statistically associated with slower completion times in some conditions for TVertMax and THorMax). In contrast, using the mouse cursor as a bookmark might be beneficial for completing TValue with Bar and Plain; yet, the evidence is not strong enough to recommend interventions regarding mouse use.

6.4 The Shape of Tables

Tables are generally more dense vertically than horizontally. This is because text, at least in the western tradition, is conventionally written left to right, making the natural shape of cells and headings rectangular and horizontally aligned. This has consequences for most tasks.

For example, we saw that after finding the maximum in THorMax, it usually only takes a single saccade (a jump) to get to the column heading, whereas in TVertMax, the last step to reach the row headers is most often a slow scan right to left. This also has implications for TValue because it requires multiple heading checks which are easier when following a row (vertical checks) than when following a column (horizontal check). In cases like these, the width of, and the spacing between, columns becomes an asset that trades off with the table's information density. We therefore recommend to **consider carefully the row-column mapping of the data** (a table and its transposed version will not usually be equivalent), and to **balance the data density of the table with people's ability to quickly navigate it**. Our decomposition of task operations is a good starting point for these considerations.

7 Limitations and Future Work

Although this work represents the broadest empirical investigation of tasks on tables to date, most are low-level tasks. Many other tasks take place on tables; tables can be of different physical sizes and data densities; and tables can contain mixed numerical and categorical/text data. Performance and behavior on those tasks will offer surprises just as those we found for our tasks. This makes results difficult to fit to existing models of graphical perception (e.g., [34]). Our data also does not enable us to validate such models because we would have to infer operations from noisy gaze movements, model the tasks in the different models, and consider the different possible strategies for each. Further work to refine models for this problem is worthy of exploration. While more precise gaze recognition could offer value in the future (e.g., to guide the design new visual aids), our analysis already provided valuable insights and explanations for performance differences based on empirical data.

We believe that the visual aids we studied are the most likely to affect reading performance; however, further research should consider the effect of more prosaic visual aids, such as separator lines and general spacing. Interaction with digital tables (e.g., interactive transient highlighting of rows or columns) also has received little attention, which is surprising given the puzzling UI and display decisions implemented in existing spreadsheet software. Our results suggest that interactivity that fits the requirements of the most important tasks is a promising source of performance improvements. Some of these optimizations could rely on empirical measurements of key parameters. For example, what is the distance and width at which readers start feeling confident with quick saccades to headings? We also believe that there is plenty of space for innovation in new methods to facilitate reading tables, from static approaches such as better versions of the Zebra and hybrid encodings to dynamic gaze-contingent dynamic techniques.

8 Conclusion

Despite the importance and ubiquity of tables, most advice about how to build them is based on experience and speculation, not empirical evidence. We carried out an experiment with four tasks that supports the use of visual encodings to save cognitive effort when finding maxima, and of zebra shading in the other tasks we tested. Visual aids on tables are more than "chart junk".

Supplemental Materials

The supplemental materials, available at https://osf.io/jfg3h/?view_only=f064cff189c4440299a3c3b10ddab232, include the following directories:

- `Preliminary task`: the description of the preliminary task and a Sankey Diagram that shows gaze behavior for this preliminary task.
- `Additional Figures`: two additional figures that represent the most typical gaze pattern for THorMax and TVertMax, and three additional figures that show the errors for THorMax, TDiff and TValue.
- `Data and Time-Error-Mouse quant analysis`: the folders necessary to reproduce the quantitative analysis of completion time, error and mouse use. The files are Jupyter notebooks that run on the R kernel (R version 4.1.3) and require the installation of JAGS (JAGS version 4.3.0) and RunJAGs as interface between the two. The data is in CSV format in the folder: `./Data and Time-Error-Mouse quant Analysis/Analysis/data/test/`.
- `Gaze Analysis`: the coded data from the visual coding of gaze videos and the description of the codes.
- `Stimuli`: the image files of the tables used in the experiment.

The pre-registration for this research is available at https://osf.io/b67xu?view_only=b9cc56507fc54ae399d0f468d53474ed.

Acknowledgments

This article is based on Yongfeng Ji's Masters thesis [24]. This research was funded in part by NSERC (2019-05422 and 2020-04401)

References

[1] R. Bachi. *Graphical rational patterns: A new approach to graphical presentation of statistics*. Transaction Publishers, 1968. 2

[2] L. Bartram, M. Correll, and M. Tory. Untidy data: The unreasonable effectiveness of tables. *IEEE Transactions on Visualization and Computer Graphics*, 28(01):686–696, jan 2022. doi: 10.1109/TVCG.2021.3114830 1

[3] J. R. Beniger and D. L. Robyn. Quantitative graphics in statistics: A brief history. *The American statistician*, 32(1):1–, 1978. doi: 10.1080/00031305.1978.10479235 2

[4] J. Bertin. *Graphics and graphic information processing*. Walter de Gruyter, 2011. 2

[5] A. Black and K. L. Stanbridge. Documents as 'critical incidents' in organization to consumer communication. *Visible language*, 46(3):246–249, 2012. 2

[6] U. Brandes, B. Nick, B. Rockstroh, and A. Steffen. Gestaltlines. *Computer graphics forum*, 32(3pt2):171–180, 2013. doi: 10.1111/cgf.12104 2

[7] C. A. Brewer, A. M. MacEachren, L. W. Pickle, and D. Herrmann. Mapping mortality: Evaluating color schemes for choropleth maps. *Annals of the Association of American Geographers*, 87(3):411–438, 1997. doi: 10.1111/1467-8306.00061 2

[8] W. C. Brinton. *Graphic methods for presenting facts*. Engineering magazine company, 1919. 2

[9] W. C. Brinton. *Graphic presentation*. Brinton associates, 1939. 2

[10] M. Campbell-Kelly. *The History of Mathematical Tables: From Sumer to Spreadsheets*. Oxford University Press, Oct. 2003. 1, 2

[11] W. S. Cleveland and R. McGill. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association*, 79(387):531–554, 1984. doi: 10.1080/01621459.1984.10478080 2, 4, 9

[12] J. Czekanowski. *Zur differentialdiagnose der neandertalgruppe*. Friedr. Vieweg & Sohn, 1909. 2

[13] C. G. Durbin. Effective Use of Tables and Figures in Abstracts, Presentations, and Papers. *Respiratory Care*, 49(10):1233–1237, Oct. 2004. 1

[14] J. Enders. Zebra striping: does it really help? In *Proceedings of the 19th Australasian conference on computer-human interaction*, OZCHI '07, pp. 319–322. ACM, 2007. doi: 10.1145/1324892.1324958 2, 3, 4, 9

[15] J. Enders, P. Clancey, L. Overkamp, P. Brosset, S. V. Prater, and M. Wills. Zebra striping: More data for the case. *A List Apart*, (267), Sep 2008. 2, 4, 9

[16] R. A. Feinberg and H. Wainer. Extracting Sunbeams From Cucumbers. *Journal of Computational and Graphical Statistics*, 20(4):793–810, Jan. 2011. doi: 10.1198/jcgs.2011.204a 1

[17] L. E. Franzblau and K. C. Chung. Graphs, tables, and figures in scientific publications: The good, the bad, and how not to be the latter. *The Journal of hand surgery (American ed.)*, 37(3):591–596, 2012. doi: 10.1016/j.jhsa.2011.12.041 1, 2

[18] K. Furmanova, S. Gratzl, H. Stitz, T. Zichner, M. Jaresova, A. Lex, and M. Streit. Taggle: Combining overview and details in tabular data visualizations. *Information Visualization*, 19(2):114–136, 2020. doi: 10.1177/1473871619878085 2

[19] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian data analysis*, vol. 2. Chapman & Hall/CRC Boca Raton, FL, USA, 2014. 4

[20] H. L. Han and M. A. Nacenta. The effect of visual and interactive representations on human performance and preference with scalar data fields. In *Proceedings of Graphics Interface 2020*, GI 2020, pp. 225 – 235. Canadian Human-Computer Communications Society / Société canadienne du dialogue humain-machine, 2020. doi: 10.20380/GI2020.23 2, 3, 9

[21] J. Hartley. Tabling information. *The American psychologist*, 46(6):655–656, 1991. doi: 10.1037/0003-066X.46.6.655 1

[22] T. Höllt, J. Beyer, F. Gschwantner, P. Muigg, H. Doleisch, G. Heinemann, and M. Hadwiger. Interactive seismic interpretation with piecewise global energy minimization. In *2011 IEEE Pacific Visualization Symposium*, pp. 59–66. IEEE, 2011. doi: 10.1109/PACIFICVIS.2011.5742373 2

[23] B. Jelen. Using sparklines to visualize your data. *Strategic finance (Montvale, N.J.)*, 92(12):62–, 2011. 2

[24] Y. Ji. *Reading Numeric Data Tables: Viewer Behavior and the Effect of Visual Aids*. Theses, University of Victoria, Dec. 2023. 10

[25] D. A. Keim, M. C. Hao, U. Dayal, and M. Hsu. Pixel bar charts: A visualization technique for very large multi-attribute data sets. *Information visualization*, 1(1):20–34, 2002. doi: 10.1057/palgrave.ivs.9500003 2

[26] J. Kruschke. *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press, 2014. 4

[27] M. Lee, T. Kent, C. M. Carswell, W. Seidelman, and M. Sublette. Zebra-striping: Visual flow in grid-based graphic design. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 58(1):1318–1322, 2014. doi: 10.1177/1541931214581275 2, 3, 9

[28] A. Lex, N. Gehlenborg, H. Strobelt, R. Vuillemot, and H. Pfister. Upset: Visualization of intersecting sets. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1983–1992, 2014. doi: 10.1109/TVCG.2014.2346248 2

[29] T. Loua. *Atlas statistique de la population de Paris*. J. Dejey & cie, 1873. 2

[30] M. MacDonald-Ross. How numbers are shown: A review of research on the presentation of quantitative data in texts. *AV communication review*, 25(4):359–409, 1977. doi: 10.1007/BF02769746 1

[31] C. Manteau, M. Nacenta, and M. Mauderer. Reading small scalar data fields: Color scales vs. detail on demand vs. fatfonts. In *Proceedings of Graphics Interface 2017*, GI 2017, pp. 50 – 56. Canadian Human-Computer Communications Society / Société canadienne du dialogue humain-machine, 2017. doi: 10.20380/GI2017.07 2, 3, 9

[32] F. T. Marchese. Exploring the Origins of Tables for Information Visualization. In *2011 15th International Conference on Information Visualisation*, pp. 395–402, July 2011. doi: 10.1109/IV.2011.36 2

[33] F. T. Marchese. Tables and Early Information Visualization. In F. T. Marchese and E. Banissi, eds., *Knowledge Visualization Currents: From Text to Art to Culture*, pp. 35–61. Springer, London, 2013. doi: 10.1007/978-1-4471-4303-1_3 1, 2

[34] J. Meyer. Performance with tables and graphs: effects of training and a Visual Search Model. *Ergonomics*, 43(11):1840–1865, Nov. 2000. doi: 10.1080/00140130050174509 9

[35] K. Moreland. Diverging color maps for scientific visualization. In *Advances in Visual Computing*, Lecture Notes in Computer Science, pp. 92–103. Springer Berlin Heidelberg, Berlin, Heidelberg. doi: 10.1007/978-3-642-10520-3_9 2

[36] P. S. Morrison. Symbolic Representation of Tabular Data. *New Zealand Journal of Geography*, 79(1):11–18, 1985. doi: 10.1111/j.0028-8292.1985.tb00199.x 1

[37] M. Nacenta, U. Hinrichs, and S. Carpendale. Fatfonts: Combining the symbolic and visual aspects of numbers. In *Proceedings of the Interna-*

tional Working Conference on Advanced Visual Interfaces, AVI '12, p. 407–414. Association for Computing Machinery, New York, NY, USA, 2012. doi: 10.1145/2254556.2254636 2

[38] C. Perin, P. Dragicevic, and J.-D. Fekete. Revisiting bertin matrices: New interactions for crafting tabular visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2082–2091, 2014. doi: 10.1109/TVCG.2014.2346279 2

[39] R. Rao and S. K. Card. The table lens: merging graphical and symbolic representations in an interactive focus + context visualization for tabular information. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '94, p. 318–322. ACM, New York, 1994. doi: 10.1145/191666.191776 2

[40] M. L. Resnick and C. Fares. Visualizations to facilitate online tabular presentation of product data. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 48(13):1498–1502, 2004. doi: 10.1177/154193120404801306 2

[41] C. Saver. Tables and figures: Adding vitality to your article. *AORN Journal*, 84(6):945–950, 2006. doi: 10.1016/S0001-2092(06)63991-4 1

[42] H. Siirtola. Bars, Pies, Doughnuts & Tables – Visualization of Proportions. In *Proceedings of the 28th International BCS Human Computer Interaction Conference (HCI 2014) (HCI)*. BCS Learning & Development, Sept. 2014. doi: 10.14236/ewic/HCI2014.38 2

[43] S. Silva, B. Sousa Santos, and J. Madeira. Using color in visualization: A survey. *Computers & graphics*, 35(2):320–333, 2011. doi: 10.1016/j.cag.2010.11.015 2

[44] R. Stauffer, G. J. Mayr, M. Dabernig, and A. Zeileis. Somewhere over the rainbow: How to make effective use of colors in meteorological visualizations. *Bulletin of the American Meteorological Society*, 96(2):203–216, 2015. doi: 10.1175/BAMS-D-13-00155.1 2

[45] E. R. Tufte. *Envisioning Information*. Graphics Pr, Cheshire, Conn, first edition ed., May 1990. 2

[46] E. R. Tufte. *Beautiful evidence*. Graphics Press, Cheshire, CT, 2006 - 2006. 1, 2

[47] H. Wainer. Improving Tabular Displays, With NAEP Tables as Examples and Inspirations. *Journal of Educational and Behavioral Statistics*, 22(1):1–30, Mar. 1997. doi: 10.3102/10769986022001001 1, 2

[48] A. P. Wheeler. Tables and graphs for monitoring temporal crime trends: Translating theory into practical crime analysis advice. *International journal of police science & management*, 18(3):159–172, 2016. doi: 10.1177/1461355716642781 1, 2

[49] L. Wilkinson and M. Friendly. The history of the cluster heat map. *The American statistician*, 63(2):179–184, 2009. doi: 10.1198/tas.2009.0033 2

[50] P. Wright and K. Fox. Presenting information in tables. *Applied Ergonomics*, 1(4):234–242, 1970. doi: 10.1016/0003-6870(70)90133-X 1, 2

[51] A. Zeileis and K. Hornik. Choosing color palettes for statistical graphics. WorkingPaper 41, Department of Statistics and Mathematics, WU Vienna University of Economics and Business, 2006. 2